PB83-193508

Confidentiality and Freedom of Information for Epidemiological Data in Governmental Research

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

1981

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/J-81-619	2. JOURNAL ARTICLE
	3. REC PB83-193508
4. TITLE AND SUBTITLE Confedentiality and Freedom of Information for Epidemiological Data in Governmental Research	5. REPORT DATE 1981
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Gerald J. Nehls, Carl G. Hayes, and William C. Nelson	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Health Effects Research Laboratory Office of Research and Development US Environmental Protection Agency Research Triangle Park, NC 27711	10. PROGRAM ELEMENT NO. AAGA1D
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Office of Research and Development Health Effects Research Laboratory US Environmental Protection Agency Research Triangle Park, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE EPA-600/11

15. SUPPLEMENTARY NOTES

Published In: Environmental Research 160-166, 1981

16. ABSTRACT

The issues of condidentiality and freedom of information are discussed for studies involving health surveillance around point sources of pollution. The problems discussed are related to protectiong data after a study has been initiated. Research studies describing the impact of confidentiality on response rates are discussed. The U.S. Environmental Protection Agency's current and proposed procedures for handling and releasing these data are presented. Theses can be used by other agencies in defining their security procedures.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS (This Report) Unclassified	21. NO. OF PAGES 8
	20. SECURITY CLASS (This page) Unclassified	22. PRICE

REPRODUCED BY NATIONAL TECHNICAL INFORMATION SERVICE U.S. DEPARTMENT OF COMMERCE SPRINGFIELD, VA. 22161

EPA-600/J-81-619
JOURNAL ARTICLE

ENVIRONMENTAL RESEARCH 25, 160–166 (1981)

Confidentiality and Freedom of Information for Epidemiological Data in Governmental Research

GERALD J. NEHLS,[1] CARL G. HAYES, AND WILLIAM C. NELSON

Health Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711

Received March 18, 1980

The issues of confidentiality and freedom of information are discussed for studies involving health surveillance around point sources of pollution. The problems discussed are related to protecting data after a study has been initiated. Research studies describing the impact of confidentiality on response rates are discussed. The U.S. Environmental Protection Agency's current and proposed procedures for handling and releasing these data are presented. These can be used by other agencies in defining their security procedures.

INTRODUCTION

The balance between confidentiality and freedom of information is a critical issue in epidemiological studies and one which will not be resolved immediately. Confidentiality can be defined as the status accorded data requiring protection from unauthorized disclosure. Freedom of Information provisions stem from legislation passed in 1967 and amended in 1974 authorizing public access to records in the possession of federal agencies. Similar state laws and policies exist pertaining to these issues.

New legislation has been introduced into Congress on these issues including a Privacy of Research Records Act affecting both the federal government and its contractors and grantees. Another bill, HR 359, has been introduced to provide for the privacy of certian records used or maintained by state and private agencies under programs receiving federal financial assistance. The U.S. Department of Justice is considering amendments to the Freedom of Information Act (FOIA). Confidentiality and freedom of information will remain an active topic for legislation for some years.

Much health surveillance is conducted by state or city health departments, frequently in collaboration with or financed by federal agencies. These departments are now and will continue to be affected by this legislation. The existing legislation has not been fully interpreted by the courts and current proposed legislation is not completely definitive, so there will be uncertainty associated with these issues for years to come. Agencies must establish policies which allow for dissemination of the studies while protecting the confidentiality of the participants. Current and proposed procedures of the U.S. Environmental Protection Agency (EPA) are presented below to illustrate the issues which must be addressed.

[1] Address requests for reprints to Gerald J. Nehls, Mail Drop 55, Data Management Branch, Biometry Division, Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711.

0013-9351/81/030160-07$02.00/0
Copyright © 1981 by Academic Press, Inc.
All rights of reproduction in any form reserved.

U.S. Government is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

SCOPE OF PROBLEM

Health surveillance around point sources of pollution involves gathering sensitive personal and socioeconomic data on potential control or covariate variables. In interview studies, these include questions on smoking history, age, race, educational level, and type of dwelling. Sensitive medical symptoms or measurements may also be collected and correlated with measurements of environmental contaminants to test hypotheses relating morbidity or mortality to pollution.

The subjects selected for the studies frequently come close to exhausting the universe of people available with given attributes thereby increasing the possibility of individual identification. The highest exposures around a pollutant point source are geographically restricted, and study designs frequently require either medically identifiable subjects (bronchitics, for example) or subjects defined institutionally for design convenience (such as primary schoolchildren or parents).

Legislation and policy are needed both to (1) protect people who participate in these studies from an unwarranted invasion of personal privacy; and (2) to assist scientists in their decision where there is ethical ambiguity between privacy and confidentiality. The area of utmost concern involves release of computerized files, or microdata. Even though obvious identifiers such as names, addresses, and telephone numbers, are deleted from the microdata, it is possible to prepare cross-tabulations of the data to isolate some individuals or families who participated in the study. Then, by comparing with other available systems of records, these individuals or families can be identified. Other systems of records such as plant employment records would also involve the same potential for personal identification.

There are two aspects to this problem: one of gaining access to sensitive data to undertake a study; the other of protecting data after a study has been initiated. The latter point is addressed because the experience and concern of the EPA is in this area.

Many other agencies are potentially affected by these issues. For example, the Occupational Safety and Health Administration (OSHA) received 212 responses to proposed regulations which would allow OSHA access to medical records of employees. Some concern was expressed that the regulations "do not contain positive guarantees that the records, once in Federal hands, would not be disclosed to others or used to make adverse decisions about individuals" (*Access Reports*, 1978b).

The National Institute of Occupational Safety and Health (NIOSH) is also affected. In a 1977 decision granting NIOSH access to medical records of a DuPont chemical plant, the court ruled that "the information not be disclosed to other governmental agencies or to anyone else under the Freedom of Information Act" (*Access Reports*, 1978a).

The Center for Disease Control, Department of Energy, Food and Drug Administration, National Cancer Institute, and the National Institute of Environmental Health Sciences are among the other agencies affected.

The issue with regard to grantees and contractors is less obvious. The Supreme Court has agreed to determine whether records of government grantees are subject to Freedom of Information requests (*Access Reports*, 1979). With contrac-

tors, information which will eventually be supplied to the government is subject to the FOIA at the time it is supplied to the government. Other applicability may depend on the wording of the funding instrument. Various state and university regulations and policies may also apply.

IMPACT OF UNAUTHORIZED DISCLOSURE

Two recent papers offer data related to the impact of confidentiality on survey response.

An experimental survey designed to measure the effect of five different levels of confidentiality statements was conducted (Goldfield *et al.*, 1977). Although the differences were not large, the trend of increasing refusal with decreasing assurance of confidentiality was statistically significant based on two nonparametic tests.

Singer (1978) also attempted to identify differences in responses to individual questions. Although it was impossible to discern any effect of confidentiality on overall response rates,

> Nonresponse rates for sensitive questions were consistently and sometimes significantly lower when people were told that their replies would be held in confidence. Furthermore there is at least a suggestion that a promise of confidentiality enhances the quality of response to the most sensitive items on the interview. Thus, if respondents can be assured of the confidentiality of their replies, response to sensitive questions will benefit. *But such assurances must be meaningful: they cannot be given lightly.* (emphasis added)

Questions on smoking, drinking, and educational level had a *significantly* higher response rate when confidentiality was pledged. These are questions asked on most environmental epidemiological surveys. Any deterioration of the response rate would have a serious impact on the quality of the study and could well influence the conclusion thereof.

Other guidance has been published in this area (Office of Federal Statistical Policy and Standards, 1978). When evaluating the impact of releasing data, examining the potential harm to the agency is recommended. A disclosure with no particularly harmful consequences to the respondent might, if highly publicized, impair an agency's ability to collect data from an increasingly distrustful public. This impairment could be particularly critical for health surveillance around point sources of pollution. This point is reinforced elsewhere (Privacy Protection Study Commission, 1977).

> Research and statistical results are too important to the common welfare to risk eroding public trust in the activities and processes that produce them.

Goldfield and his co-workers (1977) also conducted an attitude survey. In this survey, 42% of those studied said that the assurance of confidentiality made some difference in their willingness to be interviewed. Also, 40% of the people interviewed felt that the federal government was the most likely to get accurate information. The next highest organization rated 22%. Any erosion of confidence in federal studies could also be felt in nonfederal organizations. This increases the problem.

Other authors have discussed similar concerns and made suggestions (National

Academy of Sciences, 1975; Bryant and Hansen, 1975; American Statistical Association, 1977; Interagency Committee on Health Hazards of Ionizing Radiation, 1979).

EXPERIENCE AT EPA

Current precautions taken by EPA can be discussed in terms of physical security measures and information management practices. Physical measures start with entry controls to the facility. Screening devices or guards are located at all building entrances. Exterior lighting, TV cameras, and roving patrols are also used. Badges are required for entry into the facility. All data forms and files are kept under double lock. As data sets are validated, forms are shipped to the Federal Records Center operated by the General Services Administration in Atlanta. They are saved there until the records are no longer needed, usually from 5 to 15 years. Then they will be destroyed. Duplicate magnetic tape files of data sets are stored in at least two separate wings of the building, and one tape file of each data set is stored in a tape library vault.

Information management practices can best be examined by describing the steps involved in processing the data. Our current policy is that obvious identifiers such as names, addresses, and telephone numbers of participants are retained on a separate file. Linkage is available to the data file, and computer listings of the data used for validation include the names. Once validation is completed, this file is destroyed unless there is a need to retain the identifiers. They would be retained for longitudinal studies or when any additional studies are anticipated. It may be necessary to retain identifiers on a data file; however, this causes more security problems than if they were on a separate file.

Often occupational studies require sophisticated linkage over many years. If the identifiers are kept they are on a separate file and an encryption algorithm is likely to be used to code the identifiers. These algorithms are available on most computer systems.

The data reside on magnetic tapes except when actively being processed or analyzed. These tapes have file names assigned so that an unauthorized user cannot determine which data set is associated with each name. Only one person, the tape librarian, is able to identify files. The librarian does not release tapes internally for validation or analyses without first receiving approval from the supervisor. The supervisor must be aware of the reasons for releasing the data sets and must make certain the requestor is aware that the data sets contain confidential data and require special security.

Usually, the first step in the validation process is to copy the file onto a disk. This file contains a "read key" known only to the person who established the file. Should an unauthorized user determine which file contains the data he is interested in, he cannot read the file without knowing the read key.

The computer system itself offers another level of security. In order to access the system, a person must be a registered user and have a password. Without proper identification a user cannot access the computer through a terminal or with a computer program.

Listings used for validation and quality control contain confidential data. When

the process is completed, these listings are either shredded or incinerated. These listings cannot be recycled or discarded with other trash because the potential for violation of confidentiality is too great. Listings containing personal data have appropriate warnings.

One further precaution we have instituted is to have the data located on separate files and on separate tape reels rather than being in a single data base; therefore, if an unauthorized user could somehow gain access to one file, he still would not have access to the entire data base.

Another approach which can be investigated is the use of a Data Base Management System. Some of these have multiple levels of password protection. This allows the data administrator to restrict user access to specified portions of the data base. This procedure may be used in the future.

If data are released to or gathered by an EPA contractor or grantee, the funding instrument contains clauses indicating the confidentiality of the data and specifying restricted usage of these data. Contractors have copies of the data only during the time necessary to fulfill their contractual obligation to EPA. There may be cases when the contractor or grantee has some rights to the data after his period of performance with EPA. It is essential that stringent security procedures be in effect at the extramural agency before this is considered.

All reports and publications prepared from the data are checked to make certain that no individual can be identified from any tables or text.

To assure data integrity, statistical packages and customized software are used for range checks, outlier analyses, and consistency checks. Values are changed only after authorization from the epidemiologist and only after appropriate notation has been made on the input forms. After all updates and corrections have been made, a quality control process is begun with the Dodge–Romig acceptance sampling technique (Chamblee, 1979). This procedure eventually results in a file which contains an error rate less than some previously selected proportion.

There are examples of protection that could be applied to data while they are being processed or archived.

RECOMMENDATIONS FOR RELEASE OF DATA

EPA has just completed negotiations with an FOIA requestor for release of some microdata on health studies. These negotiations were undertaken after EPA refused the request, citing the FOIA exemption covering the unwarranted invasion of personal privacy. Since the requestor was sensitive to the privacy issue, the negotiations were completed successfully. As a result of this and previous experience, the following practices are now followed.

All obvious identifiers gathered in research and statistical studies involving humans do not appear in any information, tabulation, publications, or microdata released to requestors. This usually involves names, addresses, telephone numbers, and month and day of birth.

The possibility also exists that release of data can unintentionally identify individuals who participated in the study. The identification of such individuals also reveals their socioeconomic, medical, or symptom-reporting data which are of a highly sensitive nature. To protect the individuals, the following rules are applied:

Rule of three. No tabulation will be released that shows data in cells for which the count is three or less. No microdata will be released if obvious individual data can readily be tabulated so that cells containing three or fewer entries exist.

Impact of external data. When other systems of records are available to the requestor or external user on the population involved in the research studies, the rule of three may be applied to other categories prior to release of microdata. These categories are ones contained in the other system of records which would not otherwise be available to the external user. However, since there are many files, both governmental and nongovernmental, containing information about persons, there is always the possibility that cross-tabulation of the data from one file with those from another file may lead to unique categories and identify individuals. EPA cannot protect against all these possibilities but will protect against files which it has identified.

Size of population. Where the people selected for the research come close to exhausting an identifiable population, special care must be taken. EPA considers this in designing studies. If this occurs, EPA attempts to restrict its analyses to a subfile of the data which could be released. If this is not possible, EPA denies access to the data.

In many cases it is possible to design studies so that it is unnecessary to deny all access to the data, but in the class of studies called "panel studies," which follow a small number of persons over time, this is difficult.

EPA attempts to modify tabulations and microdata to satisfy the requestor by using the following techniques.

1. Combining two or more categories so that the reporting units in the resulting category are more than three; the requestor is notified as to which categories were combined.
2. Introduction of "noise" into the data, such as additional entries to certain categories or scrambling of certain data elements. Such additions are usually a less desirable alternate.
3. Creation of a subset of the data for release which reduces the possibility of identification of individuals.

Where categories must be combined, EPA attempts to do this prior to conducting its analysis, so the requestor has access to the same file used by EPA. EPA also fully describes the analysis procedure used.

EPA has implemented these practices to protect the confidentiality of individuals who participate in such studies. The data sets can be released as intended under the FOIA for further research and to organizations who may be affected by regulations which are based on the studies. These practices are recommended for the consideration of other agencies involved in collecting, processing, or analyzing such data.

REFERENCES

Access Reports (1978a). Court grants federal agency access to medical records of DuPont workers. January 10, 1978, pp. 4–5.

Access Reports (1978b). Opening worker's health files viewed as invasion of privacy. September 19, 1978, pp. 7–9.

Access Reports (1979). Supreme Court to review whether FOIA applies to records of recipient of government grants. May 30, 1979, p. 6.

American Statistical Association (1977). Report of ad hoc committee on privacy and confidentiality. *Amer. Statist.* **31**, 59–78.

Bryant, E. C., and Hansen, M. H. (1975). Invasion of privacy and surveys: A growing dilemma. *In* "Perspective on Attitude Assessment. Surveys and their Alternatives" (H. W. Sinaiko and L. A. Broedling, Eds.), pp. 75–86. Smithsonian Institute, Washington, D.C.

Chamblee, C. (1979). Use of statistical sampling in validating health effects data. *In* "Proceedings of the Environmental Protection Agency Data Validation Conference," in press.

Goldfield, E. D., Turner, A. G., Cowan, C. D., and Scott, J. C. (1977). "Privacy and Confidentiality as Factors in Survey Responses." Presented at 1977 Annual Meeting of the American Statistical Association.

Interagency Committee on Health Hazards of Ionizing Radiation (1979). "Report of the Work Group on Records and Privacy." U.S. Department of Health, Education, and Welfare, Washington, D.C.

National Academy of Sciences (1975). "Protecting Individual Privacy in Evaluation Research." Nat. Acad. Sci. Washington, D.C.

Office of Federal Statistical Policy and Standards (1978). "Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques." U.S. Govt. Printing Office, Washington, D.C.

Privacy Protection Study Commission (1977). "Personal Privacy in an Information Society." U.S. Govt. Printing Office, Washington, D.C.

Singer, E. (1978). Informed consent: Consequences for response rate and response quality in social surveys. *Amer. Sociol. Rev.* **43**, 144–162.